End-to-end Generative Pretraining for Multimodal Video Captioning

Year: 2022 | Citations: 184 | Authors: P. H. Seo, Arsha Nagrani, Anurag Arnab, C. Schmid

Abstract

Recent video and language pretraining frameworks lack the ability to generate sentences. We present Multimodal Video Generative Pretraining (MV-GPT), a new pretraining framework for learning from unlabelled videos which can be effectively used for generative tasks such as multimodal video captioning. Unlike recent video-language pretraining frameworks, our framework trains both a multimodal video encoder and a sentence decoder jointly. To overcome the lack of captions in unlabelled videos, we leverage the future utterance as an additional text source and propose a bidirectional generation objective - we generate future utterances given the present mulitmodal context, and also the present utterance given future observations. With this objective, we train an encoder-decoder model end-to-end to generate a caption from raw pixels and transcribed speech directly. Our model achieves state-of the-art performance for multimodal video captioning on four standard benchmarks, as well as for other video understanding tasks such as VideoQA, video retrieval and action classification.